Case-Based Reasoning for Financial Prediction

Dragan Simić[*], Zoran Budimac[**], Vladimir Kurbalija[**], Mirjana Ivanović[**]

[*]Novi Sad Fair, Hajduk Veljkova 11, 21000 Novi Sad, Serbia & Montenegro
dsimic@nsfair.co.yu

[**]Department of Mathematics and Informatics, Faculty of Science,
University of Novi Sad,
Trg D. Obradovića 4, 21000 Novi Sad, Serbia & Montenegro
{zjb,kurba,mira}@im.ns.ac.yu

Abstract. A concept of financial prediction system is considered in this paper. By integrating multidimensional data technology (data warehouse, OLAP) and case-based reasoning, we are able to predict financial trends and provide enough data for business decision making. Methodology has been successfully used and tested in the management information system of "Novi Sad Fair".

1 Introduction

A data warehouse can be defined as an online repository of historical enterprise data and is optimized for complex data analysis operations rather than transactional processing. OLAP (online analytical processing) refers to technologies that allow users to efficiently retrieve and aggregate data along the dimensions with different functions from the data warehouse. In order to help knowledge workers (executives, managers, analysts), to focus on important data, and to make better decisions, case-based reasoning (CBR - an artificial intelligence technique) is introduced for making predictions based on previous cases. CBR will automatically generate an answer to the problem using stored experience, thus freeing the human expert of obligations to analyze numerical or graphical data.

The use of CBR in predicting the rhythm of issuing invoices and receiving actual payments based on the experience stored in the data warehouse is presented in this paper. Predictions obtained in this manner are important for future planning of a company such is the ”Novi Sad Fair”, because measures of success in business are: achievement of sales plans, revenue, and company liquidation.

The combination of CBR and data warehousing, i.e. making an OLAP intelligent by the use of CBR is a rarely used approach. The system also uses a novel CBR technique to compare graphical representation of data, which greatly simplifies the explanation of the prediction process to the end-user [2].

Performed simulations show that predictions made by CBR differ only for 8% in respect to what actually happened. With inclusion of more historical data in the warehouse, the system gets better in predictions. Although achieved results of this financial prediction present a significantly good outcome, the research on this project can be continued. There are several important issues that the research could focus on in the future.

The rest of the paper is organized as follows. The following section elaborates the original algorithm for searching the previous cases (curves) looking for the most similar one. Section three describes the actual application of our technique to the given problem for inclusion of CBR in decision support system, while section four describes results and measurements. Fifth section elaborates several important issues which could improve the existing system. Section six presents the related work, while the seventh section concludes the paper.

2 CBR for Predicting Behavior of Curves

Generally speaking, case-based reasoning is applied to solving new problems by adapting solutions that worked for similar problems in the past. The main supposition here is that similar problems have similar solutions. The basic scenario for mainly all CBR applications looks as follows: in order to find a solution of an actual problem, one looks for a similar problem in an experience base, takes the solution from the past and uses it as a starting point to find a solution to the actual problem. Experience is stored in a form of cases in CBR systems. The case is a recorded situation where problem was totally or partially solved, and it can be represented as an ordered pair *(problem, solution)*. The whole experience is stored in *case base*, which is a set of cases, each case representing some previous episode where the problem was successfully solved.

The main problem in CBR is to find a good similarity measure – the measure that can tell to what extent the two problems are similar. Similarity can be defined as a function $sim : U \times CB \rightarrow [0\ ,\ 1]$ where U refers to the universe of all objects (from a given domain), while CB refers to the case base (just those objects which were examined in the past and saved in the case memory). The higher value of the similarity function means that these objects are more similar.

The main advantage of this technology is that it can be applied to almost any domain. CBR system does not try to find rules between parameters of the problem; it just tries to find similar problems (from the past) and to use solutions of the similar problems as a solution of an actual problem. So, this approach is extremely suitable for less examined domains – for domains where rules and connections between parameters are not known. The second very important advantage is that CBR approach to learning and problem solving is very similar to human cognitive processes – people take into account and use past experiences to make future decisions.

Because the aim of the system is to help end-users (executives, managers) to make business decisions we used the CBR system that makes predictions based on curves [2]. The system analyses curves, compares them to similar curves from the past and predicts the future behavior of the current curve on the basis of the most similar curves from the past.

The main problem here, as almost in every CBR system, was to create a good similarity measure for curves, i.e. what is the function that can tell to what extent the two curves are similar. In order to present two curves, it is possible to make a choice between a classical interpolating polynomial and a cubic spline. The cubic spline was chosen for two main reasons:

Power: for the n+1 points classical interpolating polynomial has the power n, while cubic spline always has the power of 4.

Oscillation: if only one point is moved (which can be the result of bad experiment or measuring) classical interpolating polynomial significantly changes (oscillates), while cubic spline only changes locally (which is more appropriate for real world domains).

When the cubic spline is calculated for curves, then one very intuitive and simple similarity measure can be used. Distance is the dual notion for similarity ($sim = 1/(1+dis)$). The distance between two curves can be represented as a surface between these curves. This surface can be easily calculated using the definitive integral. Furthermore, the calculation of the definitive integral for polynomials is a very simple and efficient operation.

3 Application of the System

A data warehouse of ”Novi Sad Fair” contains data about payment and invoicing processes from the last 4 years for every exhibition - containing between 25 and 30 exhibitions every year. Processes are presented as sets of points where every point is given with the time of the measuring (day from the beginning of the process) and the value of payment or invoicing on that day. It can be concluded that these processes can be represented as curves.

The measurement of the payment and invoicing values was done every 4 days from the beginning of the invoice process in duration of 400 days - therefore every curve consists of approximately 100 points. By analyzing these curves, the process of invoicing usually starts several months before the exhibition and that value of invoicing rapidly grows approximately to the time of the beginning of exhibition. After that time the value of invoicing remains approximately the same till the end of the process. That moment, when the value of invoicing reaches some constant value and stays the same to the end, is called the *time of saturation for the invoicing process*, and the corresponding value – the *value of saturation*.

The process of payment starts several days after the corresponding process of invoicing (process of payment and invoicing for the same exhibition). After that the value of payment grows, but not so rapidly as the value of invoicing. At the moment of exhibition the value of payment is between 30% and 50% of the value of invoicing. Then the value of payment continues to grow to some moment when it reaches a constant value and stays approximately constant till the end of the process. That moment is called the *time of saturation for the payment process*, and the corresponding value – the *value of saturation.*

Payment time of saturation is usually several months after the *invoice time of saturation*, and the *payment value of saturation* is always less than the *invoice value of saturation* or equal to it. The analysis shows that payment value of saturation is between 80% and 100% of the invoice value of saturation. The maximum represents a total of services invoiced and that amount is to be paid. The same stands for the invoice curve where the maximum amount of payment represents the amount of payment by regular means. The rest will be paid later by the court order, other special business agreements or, perhaps, will not be paid at all (debtor bankruptcy).

One characteristic invoice and a corresponding payment curve as the "Old payment curve" and "Old invoice curve" from the “curve base” are shown (Fig. 1). The points of saturation (time and value) are represented with the emphasized points on curves.

System first reads the input data from two data marts: one data mart contains the information about all invoice processes for every exhibition in the past 4 years, while the other data mart contains the information about the corresponding payment processes. After that, the system creates splines for every curve (invoice and payment) and internally stores the curves in the list of pairs containing the invoice curve and the corresponding payment curve.

In the same way system reads the problem curves from the third data mart. The problem is invoice and a corresponding problem curve at the moment of the exhibition. At that moment, the invoice curve reaches its saturation point, while the payment curve is still far away from its saturation point. These curves are shown as the “Actual payment curve” and the “Actual invoice curve” (Fig. 2).

The solution to this problem would be the saturation point for the payment curve, and a detailed calculation is shown [6]. Time point and the amount of payment of a debt are marked on the graphic by a big dot (Fig. 2).

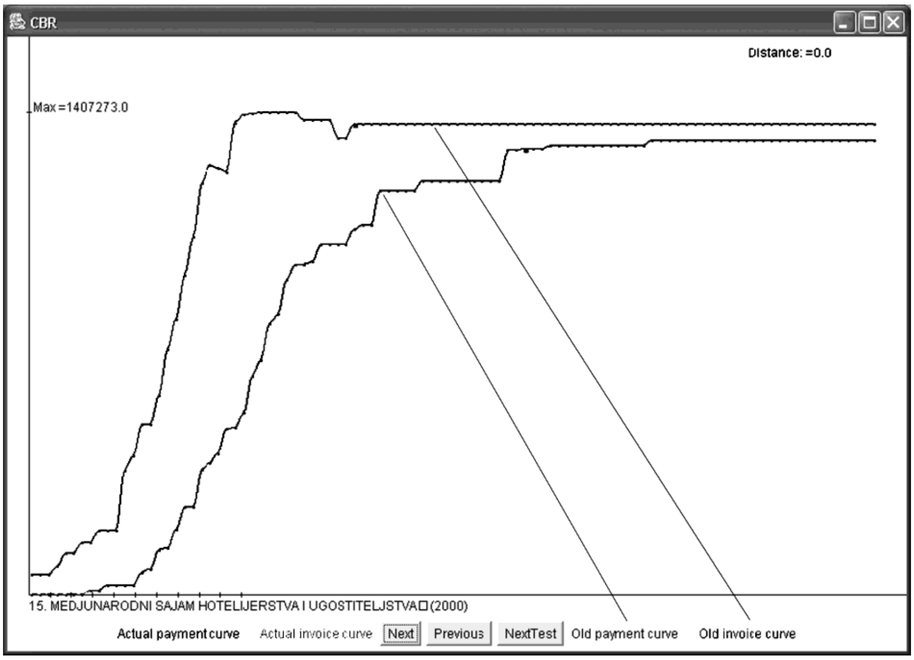


Fig. 1. The curves from data mart as the “old payment curve” and the ”old invoice curve”

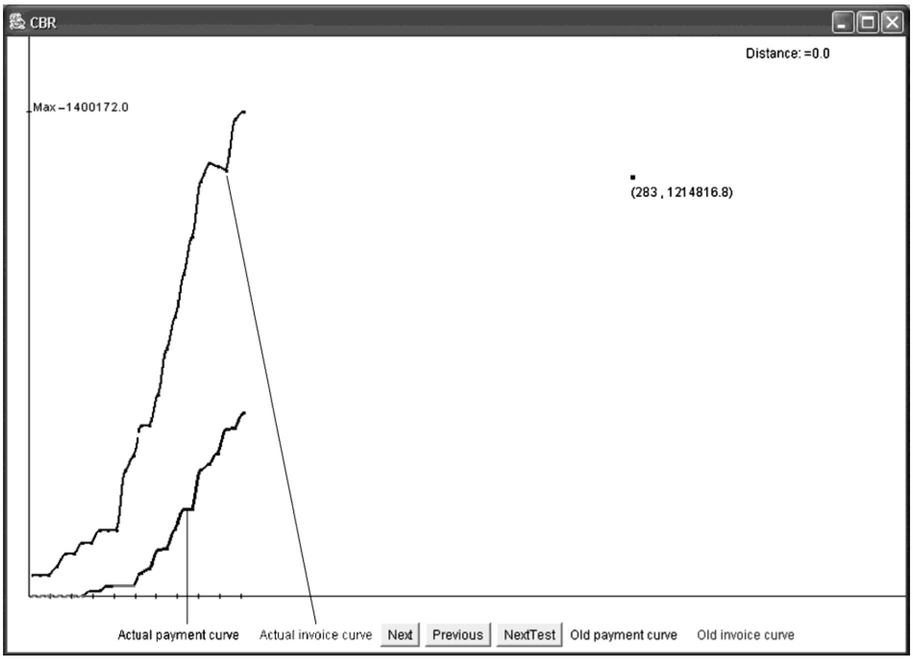


Fig. 2. Problem payment and invoice curves, as the ”Actual payment curve” and the ”Actual invoice curve” and prediction for the future payments The curves from data mart as the “old payment curve” and the ”old invoice curve”

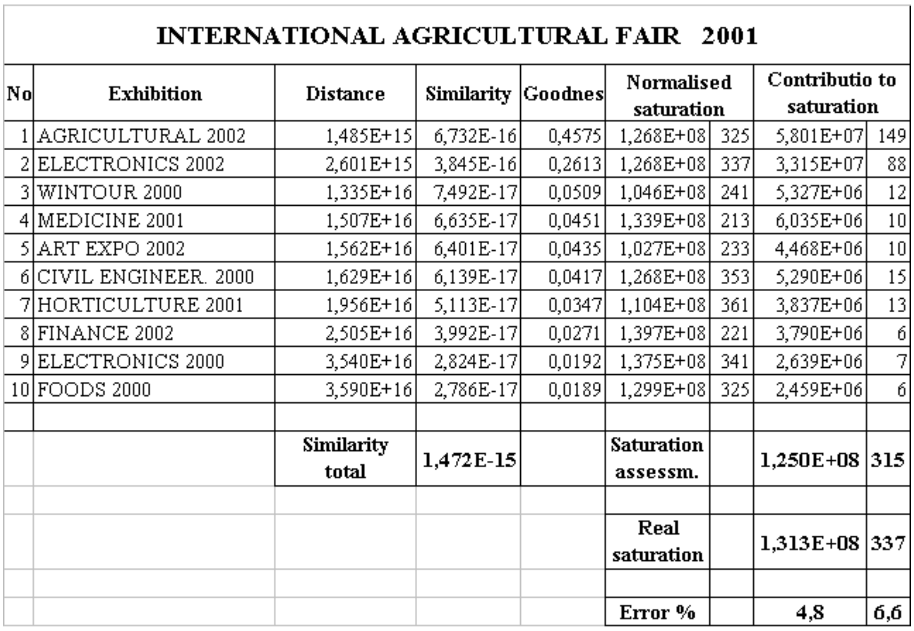
This means that system helps experts by suggesting and predicting the level of future payments. At the end of the total invoicing for selected fair exposition, operational exposition manager can get a prediction from CBR system of:

1. the time period when payment of a debt will be made
2. the amount paid regularly.

4 Results and Measurements

The measurements of already known invoice and payment value sets and payments for a number of past fair exhibitions have been made with the aim of proving accuracy of the recommended financial prediction support system for assessing the values and the time of regular payments.

The results of several conducted measurements for a number of past fair exhibitions will be shown in the following text. Measurements for the largest and the worthiest fair exhibition will be presented (“International Agricultural Fair 2001”) (Figs. 3 and 4). “International Autumn Fair 2002” measurement results will be shown as well (Figs. 5 and 6).



INTERNATIONAL AGRICULTURAL FAIR 2001

No	Exhibition	Distance	Similarity	Goodnes	Normalised saturation		Contributio to saturation	
1	AGRICULTURAL 2002	1,485E+15	6,732E-16	0,4575	1,268E+08	325	5,801E+07	149
2	ELECTRONICS 2002	2,601E+15	3,845E-16	0,2613	1,268E+08	337	3,315E+07	88
3	WINTOUR 2000	1,335E+16	7,492E-17	0,0509	1,046E+08	241	5,327E+06	12
4	MEDICINE 2001	1,507E+16	6,635E-17	0,0451	1,339E+08	213	6,035E+06	10
5	ART EXPO 2002	1,562E+16	6,401E-17	0,0435	1,027E+08	233	4,468E+06	10
6	CIVIL ENGINEER. 2000	1,629E+16	6,139E-17	0,0417	1,268E+08	353	5,290E+06	15
7	HORTICULTURE 2001	1,956E+16	5,113E-17	0,0347	1,104E+08	361	3,837E+06	13
8	FINANCE 2002	2,505E+16	3,992E-17	0,0271	1,397E+08	221	3,790E+06	6
9	ELECTRONICS 2000	3,540E+16	2,824E-17	0,0192	1,375E+08	341	2,639E+06	7
10	FOODS 2000	3,590E+16	2,786E-17	0,0189	1,299E+08	325	2,459E+06	6
		Similarity total	1,472E-15		Saturation assessm.		1,250E+08	315
					Real saturation		1,313E+08	337
					Error %		4,8	6,6

Fig. 3. Measurement results for "International Agricultural Fair 2001"

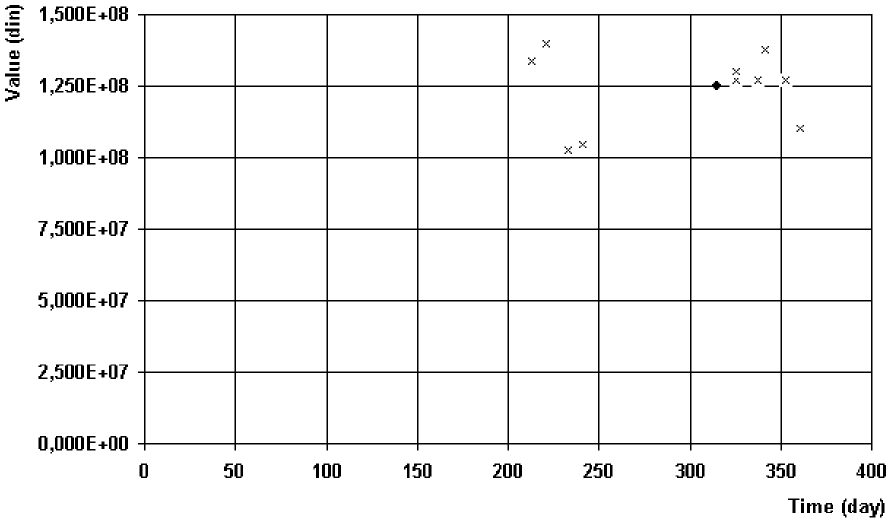
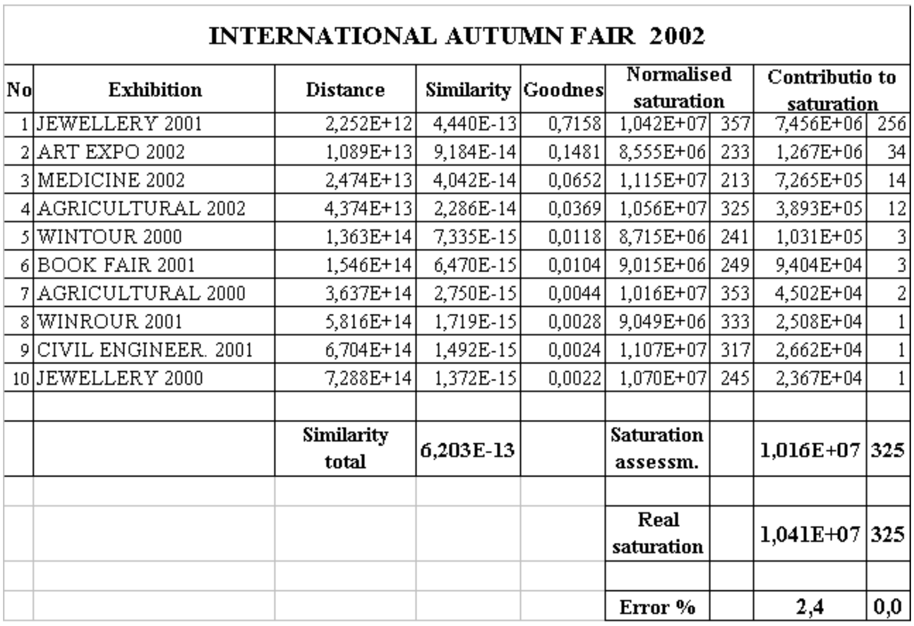


Fig. 4. Saturation assessment value and normalized saturation values (10 most similar) for "International Agricultural Fair 2001"

As it has already been stated the results of these measurements and case-based reasoning system predictions show good results with the prediction error for:

1. the value of regular debt payment from 2,4 % to 4,8 %

2. the time span of regular payment from 0,0 % to 6 %.



INTERNATIONAL AUTUMN FAIR 2002									
No	Exhibition	Distance	Similarity	Goodnes	Normalised saturation		Contributio to saturation		
1	JEWELLERY 2001	2,252E+12	4,440E-13	0,7158	1,042E+07	357	7,456E+06	256
2	ART EXPO 2002	1,089E+13	9,184E-14	0,1481	8,555E+06	233	1,267E+06	34
3	MEDICINE 2002	2,474E+13	4,042E-14	0,0652	1,115E+07	213	7,265E+05	14
4	AGRICULTURAL 2002	4,374E+13	2,286E-14	0,0369	1,056E+07	325	3,893E+05	12
5	WINTOUR 2000	1,363E+14	7,335E-15	0,0118	8,715E+06	241	1,031E+05	3
6	BOOK FAIR 2001	1,546E+14	6,470E-15	0,0104	9,015E+06	249	9,404E+04	3
7	AGRICULTURAL 2000	3,637E+14	2,750E-15	0,0044	1,016E+07	353	4,502E+04	2
8	WINROUR 2001	5,816E+14	1,719E-15	0,0028	9,049E+06	333	2,508E+04	1
9	CIVIL ENGINEER. 2001	6,704E+14	1,492E-15	0,0024	1,107E+07	317	2,662E+04	1
10	JEWELLERY 2000	7,288E+14	1,372E-15	0,0022	1,070E+07	245	2,367E+04	1
		Similarity total	6,203E-13		Saturation assessm.		1,016E+07	325
					Real saturation		1,041E+07	325
					Error %		2,4	0,0

Fig. 5. Measurement results for “International Autumn Fair 2002”

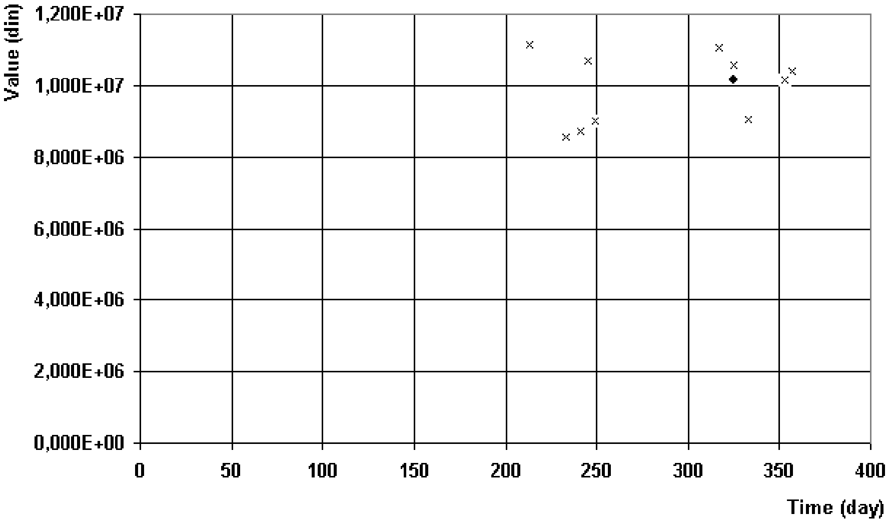


Fig. 6. Saturation assessment value and normalized saturation values (10 most similar) for ”International Autumn Fair 2002”

Other measurements have also been completed showing that when the recommended system is used with a set of already known values, financial prediction system based on CBR technique gets results which differ up to 8 % in value of what actually happened [5].

5 Improving the Existing System

Although achieved results of this financial prediction and decision support system present a significantly good outcome, the research on this project can be continued. There are several important issues that the research could focus on in the future.

5.1 Financial prediction algorithm improvement

The existing algorithm operates in a way where prediction is done at the moment of completing the fair service invoicing process. A new algorithm could provide predictions during the whole period of invoicing and payment. This means that payment prediction can be done immediately after the first invoiced service payment. Furthermore, when the invoicing reaches its saturation point and payment continues to grow towards saturation, prediction should give better results since the payment system converges to the final value of the amount of money and the payment time.

5.2 Cubic spline or linear interpolation

The other issue that needs to be looked at is the behavior of the new algorithm in case of cubic spline interpolation and linear interpolation. It is obvious that the financial prediction system depends on the usage area so that everything can be confirmed only by measuring in the domain of already known results.

5.3 Solution revision and retaining

The system has to support the revision (repeated control) of the solution and the retaining of the solution fulfilling the basic CBR model (retrieve, reuse, revise and retain). By memorizing:

a) the problem,
b) suggested solution,
c) the number of similar curves used for obtaining the suggestion, and
d) the real solution

the system uses this information in the phase of reusing the solution for future problems.

The system will then use not only 10 of the most similar curves but will also inspect the previous decisions in order to find a ‘more precise’ number of similar curves that would lead to the better prediction.

5.4 Application of other artificial intelligence techniques

It is also significant to consider the application of other artificial intelligence techniques for financial prediction support. Financial prediction as a fuzzy decision process is especially interesting for observation where its expected value benefit is established for different alternatives of stochastic dominant curve occurrence. This means that two artificial intelligence techniques would be integrated for financial prediction: case-based reasoning and fuzzy logic, as well as data warehouse information technology which would be an extremely complex model.

Using these comprehensive improvements, the theoretical basis for financial prediction and decision support would be greatly extended. However, this complex model would require a lot of effort to be transferred into a practical model.

6 Related Work

CBR is occasionally used on multidimensional technology [1, 3, 4]. [3] theoretically initiates and introduces this complex integration type (online analytical processing ↔ case-based reasoning). The system implements an environment in which automated test procedures are carried out frequently.

[4] examines the integrated system, “passive” nature of interactive CBR with an active database system. The two-layer “ActiveCBR” architecture support with active rules can perform event detecting, condition monitoring and event handling in an automatic manner.

[1] introduces CALIBRE (*Ca*ndidate *Lib*rary *Re*trieval) software tool for extraction of candidate list for target marketing campaigns. An implemented case-based reasoning combined with *data mining* as an innovative form of knowledge management for supporting this goal.

However, none of these systems can be used for financial business systems nor resembles our suggested original graphic algorithms.

7 Conclusion

The paper has in greater detail described the CBR part of the system giving a thorough explanation of one case study. One part of the paper presented the decision support system that uses CBR as an OLAP to the data warehouse. The other part introduced system improvements.

There are numerous advantages of this system. For instance, operational managers can make important business activities based on CBR predictions. They would: a) make payment delays shorter, b) make the total of payment amount bigger, c) secure payment guarantee on time, d) reduce the risk of payment cancellation, and e) inform senior managers on time. By combining graphical representation of predicted values with most similar curves from the past, the system enables better and more focused understanding of predictions with respect to real data from the past.

It is also important to consider the application of other artificial intelligence techniques for financial prediction support. Financial prediction as a fuzzy decision process is especially interesting for observation, expecting to benefit from different alternatives of stochastic dominant curve occurrence.

Presented system is not only limited to this case-study but it can be applied to other business values as well (expenses, investments, and profit) and it guarantees the same or even better level of success.

References

1. Fagan, M., Bloor, K.: Case-Based Reasoning for Candidate List Extraction in a Marketing Domain, Proceedings of 3th International Conference on Case-Based Reasoning, ICCBR-99, Seeon Monastery, Germany, Springer-Verlag: 426--437, 1999.
2. Kurbalija, V.: On Similarity of Curves – Project Report, Humboldt University, AI Lab, Berlin, 2003.
3. Shuster, A., Sterrit, R., Adamson, K., Shapcott, M., Curran, E., Case-based reasoning for complex telecommunication systems.
4. Li, S., Yang, Q.: Activating Case-Based Reasoning with Active Databases, Proceedings of 5th European Workshop, EWCBR 2000, Trento, Italy, Springer-Verlag: 3-14, 2000.
5. Simić, D., Financial Prediction and Decision Support System Based on Artificial Intelligence Technology, Ph.D. thesis, draft text, Novi Sad 2003.
6. Simić, D., Kurbalija, V., Budimac, Z., An Application of Case-Based Reasoning in Multidimensional Database Architecture, DaWaK 2003, Proceedings: 5th International Conference on Data Warehousing and Knowledge Discovery, Springer Verlag, 66-75, 2003.